Phonetic cues for phrase boundaries and dicourse finality

Cecilia Hemming
Department of Languages, University College of Skövde
Swedish National Graduate School of Language Technology

Abstract

Spoken language interfaces for real-world systems are now becoming a practical possibility. It has become apparent that such interfaces will need to gain knowledge from a variety of cues from diverse sources to be enough robust and natural. This paper summarizes some important topics of prosodic segmentation problems in automatic recognition of continuous speech in English and Swedish. It does by no means attempt to be an exhaustive presentation of the area.

Introduction

Background

The grammar of spoken utterances shares many features with the grammar of written sentences but the two approaches also differ in numerous ways. In English, spoken utterances:

- are often shorter (many single clauses)
- contain pronouns rather than full lexical noun phrases
- show syntactic disfluences (like *filled pauses* [uh, um…], *word repetitions*, *restarts* of what have been said and *repairs*, correcting something that went wrong)

Furthermore, word- and syllable boundaries do not always coincide in fluent speech. Co-articulation often affects word boundaries by deleting or changing sounds. For example: a quick pronounciation of *good boy* gives as result that the *d* in *good* is assimilated to the *b* in *boy* (McTear, 2001). Gold uses the following example to illustrate the latter problem (Gold, 2000): The word *five* in isolation can be described as having one syllable [fɑʸv], but when it is pronounced together with the word *eight* there can be a *resyllabification*. This means that it can be pronounced in such a way that *five* now also influence part of the following word's syllables [fɑʸ] [vet]. These and other spoken language features like different dialect characteristics, speaker variability and disturbing background noises make the prosodic segmentation of discourse parts a very complex task. Nevertheless, it is important to exploit low-level robust cues that are fast to process, so that slower and more complex analysis can be reserved for those inputs that require it. This is especially important in systems expected to participate in real-time conversational interaction.

Purpose

The goal of this paper is to present an overview of some facts, ideas and findings in the field of phonetic boundary cues that can have implications for speech recognition. This in order to fulfill a part of the course Speech Technology 1 within GSLT (the Swedish National Graduate School for Language Technology).

Method

I first give a short background of the topic followed by an explanation on how some researchers look at the prosodic segmentation problem of continuous speech. Then follows a presentation of different research topics that I have found in the literature. The final part contains an overall summary.

Units of analysis

A discourse text is often divided into units of some type before it is analysed. Basic units in the written language are rather clear: punctuation marks or indentation, and usually also syntactically defined: clause, sentence, paragraph. Spoken language units are of different type. Many utterances are short (see introduction, above), they are often syntactically incomplete without explicit subject or verb. Discourse researchers instead employ units based on intonation or bounded by pauses. Most phoneticians agree that speech production is best described in terms of phrase like units. Different researchers use different terms to define these: e.g. breath groups (Couper-Kuhlen, according to Edwards, 1993), idea units (Chafe, according to Edwards, 1993), informational or intonational phrases. An *intonation unit* can be defined as “a stretch of speech uttered under a single intonation contour” and tends to be marked by specific cues:

> - a pause and a shift upwards in the pitch level in the beginning
> - a lengthening of the final syllable, ending the unit
>
> (Du Bois, J.W. 1991)

A more general determination of phrase boundaries depends on what divisions are important according to rhythmic and thematic organization of the actual discourse. Phrase boundaries often serve to indicate a syntactic boundary, like the beginning of a relative clause, or a change in topic, which for the comprehension is important to detect. A phrase or a sentence is not a simple string of separate phonemes or words. Words within a sentence have mostly shorter duration and are subject to coarticulation at word boundaries. There can be phonetic changes like that seen in the broadband spectrogram below: the /t/ in *shirt* is not pronounced as a *t* in this context, it is reduced to a flap.

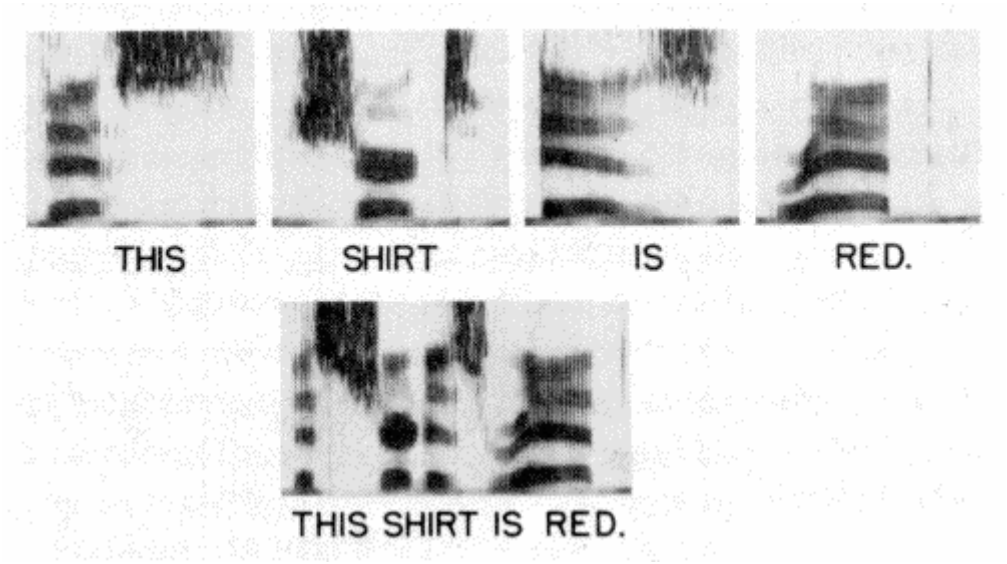


Fig. 1
Broadband spectrograms indicating that a sentence is very different from a concatenated string of words recorded in isolation (Klatt, 1987).

Acoustic values

Prosodic details help the listener segment the acoustic stream. In the picture below one can see three typical clause-final intonation contours (at the top) and an example of a fundamental frequency "hat pattern" of rises and falls between the “brim and top of a hat” for a two-clause sentence (at the bottom).

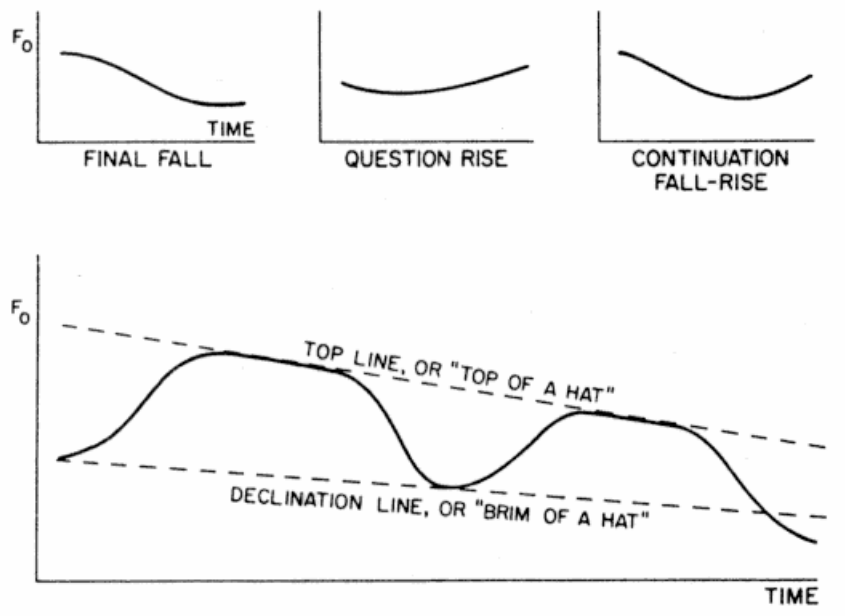


Fig 2.
Clause-final intonation contours (top), and a fundamental frequency "hat pattern" of rises and falls for a two-clause sentence (bottom). (Klatt, 1987)

The *intensity*, the *duration*, and the *fundamental frequency* (f_0) characterize a tone in physical terms. They induce the sensations of *loudness*, *length*, and *pitch*, respectively. The change over time of these parameters (intensity, duration, and f_0) can carry linguistically significant prosodic information in speech (Klatt 1989).

Physical Quantity	Nearest Subjective Attributes
Intensity pattern	syllabic structure, vocal effort, stress
Duration pattern	speaking rate, rhythm, stress, emphasis, syntactic structure
f_0 pattern	intonation, stress, emphasis, gender, vocal tract length, psychological state, attitude

Physical and subjective components of sentence prosody according to Klatt (Klatt, 1987)

Nakajima and Allen examined the relationship between fundamental frequency and discourse structure in spontaneous task-oriented dialogue and found that f_0 values tend to signal topic shift and topic continuation across pause boundaries (Nakajima & Allen, 1993). However, the correlations between acoustically measured values and linguistically significant categories are complex and far from perfect. Contours perceived as functionally equivalent, e.g. rising intonation, can have big variations acoustically: with different number of pitch peaks, varying speeds of pitch change or stretching over different lengths of speech. There is more variation in the f_0 curve than speakers intend as speech melody or listeners perceive as significant. Studies have shown that relatively big changes in f_0 can be ignored by listeners and still very small changes can produce significant differences in perceived pitch. This is due to influences like *vowel height* and *phonetic context*. Vowel quality can interfere with fundamental frequency for instance in that a vowel articulated high in the mouth have much higher f_0 and also tend to have less intensity compared with lower vowels. The place of articulation can also influence the duration both for vowels and consonants: high vowels are generally shorter than low vowels and alveolar as well as velar consonants have been found to be intrinsically shorter than labials. The surrounding sounds are also important and can influence the fundamental frequency of a sound segment. A vowel after a voiceless fricative, for example, has a higher f_0 on the average than what the same vowel pronounced after a voiced fricative has (Edwards, 1993).

Final lengthening

Speech rate tends to speed up at the beginning of phrases and slow down at ends (Edwards, 1993). Studies have shown that Final Lengthening (FL) can signal boundaries in spoken language. The domain where FL is believed to occur is the rhyme of the final syllable (Crystal & House, 1990). Is there a clear correlation between type of boundary and the phonetic correlates? Horne et al. studied radio broadcasts on Stock-Market rates to see how FL behaves as a parameter of prosodic boundary strength in Swedish (Horne et al. 1995).

In the study the Swedish word “procent” was measured when followed by four types of boundary: 0- boundary, Prosodic Word- (PW), Prosodic Phrase- (PPh) and full prosodic utterance-boundaries (PU). Moreover, 2 subcategories of boundary within the PPh and PU were distinguished: clausefinal/sentence-final position for PPh and paragraph-final/textfinal position for PU. Thus the test word was followed by 6 different boundary categories:0, PW, PPh/C, PPh/S, PU/P and PU/T, example below:

Vid 13-tiden noterades Stockholmsfondbörs generalindex till 1026,1. Det är en uppgång med 0,1 ***procent*** *(PW) jämförtmed gårdagens slutindex. 16-i-topp- index hade då gått upp med 0,4* ***procent***

(PU). Marknadsräntorna vid middagstid: den 4-åriga standardobligationen låg då stilla på gårdagens sluträntа på 10,12 ***procent*** *(PPh), 12-månaders statsskuldväxlar hade…*

All 6 categories occurred after a focused and a non focused test word respectively ([+focus] and [-focus]).

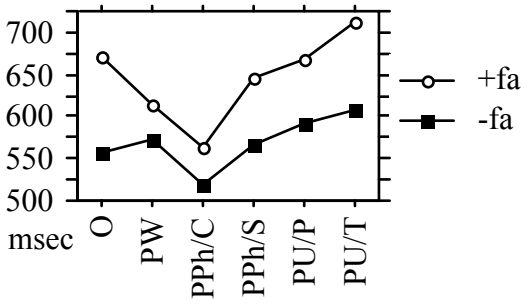


Fig. 3

Duration of test word with and without focus accent (+/- fa) before each boundary. (C=end of clause, S= end of sentence, P=end of paragraph, T=end of text) (Horne et al. 1995).

The figure shows complex effects of boundary type on word duration: an increase from PPh/C to PU/T but a decrease from 0 to PPh/C. When looking at the separate phonemes of the final syllable, it was found that indeed the FL (from the PPh level and stronger, that is to the right in the figure) exists independently of focus accent, but not on all segments. This can explain earlier disagreement in the matter.

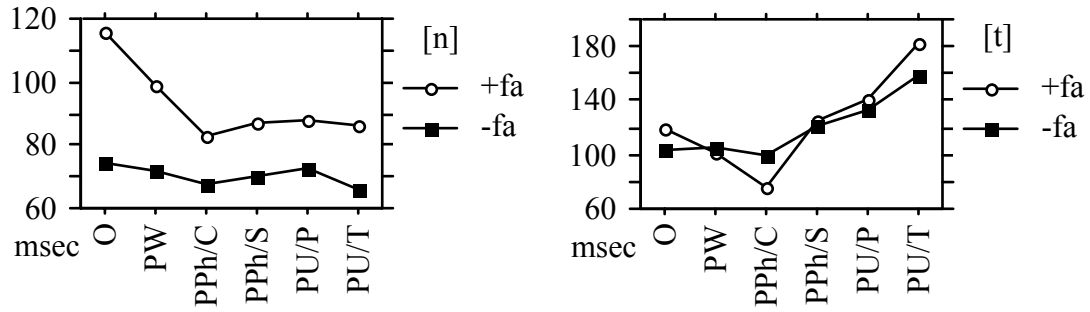


Fig. 4

Stressed final syllable segment duration ([n] to the left and [t]to the right - in the word "procent") in +/- focus accented test word before boundary (Horne et al. 1995).

Whether the test word, "procent" [prusɛnt], was focused or not had significant effects on all segments except the final [t], see figure 4. above, to the right. The increase of duration seen in fig. 3 stems primarily from the [t], while the decrease in the left side of the curve is a combined effect of adjustments made in [ɛ], [n] and also to some extent in [t]. The figure also clearly shows what Berkovits call progressive lengthening, i.e. that the lengthening regards the final consonant more than the preceding vowel (Berkovits, 1994).

Silent Interval duration

Several studies have found that pauses are used for planning and therefore often are found at clause boundaries (Chafe, in Edwards, 1993). Even if a pause is experienced as a period of nonphonation, its perceived length depends on factors as articulation rate, acoustic environment and position in the utterance. Norms vary of

course within and across cultures. British speakers may expect and use longer interturn pauses than many American speakers (Tannen in Edwards, 1993). Horne and her colleagues (see above) also investigated the relation in Swedish between FL and Silent Interval duration, SI, for which researchers earlier had found a negative correlation when studying read prose (Fant & Kruckenberg, 1989). The figure below shows the SI duration after the test word positioned in different boundary types:

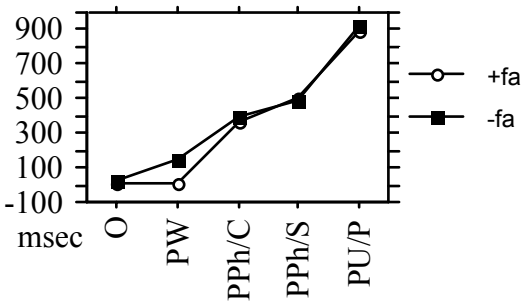


Fig. 5
Silent interval duration following the test word ("procent") with +/- focus accent (Horne et al. 1995).

One can see that the SI, just as the final consonant [t], is not affected by the +/- focus distinction. The FL-measurements in the same study, showed that the duration was greatest in the 0-boundary, less at the end of a PW and least at the end of a PPh/C word. This corroborates other observations of a trading relation between SI and FL for data below the PPh/C level.

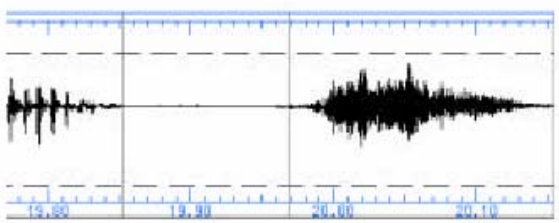
Final Aspiration

When talking about segmental *strength* and *weakness* one usually relates sound categories to a scale of sonority, see below.

Affricates/Aspirated stops **>** *Stops* **>** *Fricatives* **>** *Nasals* **>** *Liquids* **>** *Glides* **>** *Vowels*

Sonority

As seen on the scale, vowels are here assumed to be the most sonorant segment and aspirated stops or affricates the least sonorant segment. In many Germanic languages aspirated and nonaspirated voiceless stops are common variants in phonetic context. In English for instance, aspirated stops are observed to occur at the beginning of stressed syllables. This is also the case in Swedish but here they are also found in final or prepausal position (Johansson et al. 2001). An aspirated voiceless stop is a period of noisy airflow after the burst, and it lasts for a considerable period of time, see figure 6.



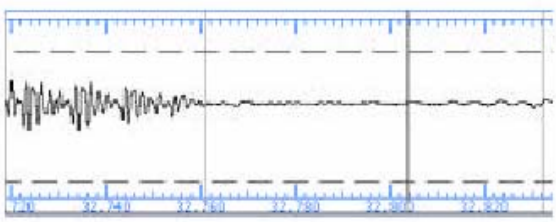


Fig. 6
Waveform of "att" with final aspiration *Waveform of "att" without final aspiration (Johansson et al. 2001).*

Johansson and her colleagues describe the aspiration as a kind of final lengthening that involves the release phase of the final stop consonant. They observe a correlation between pauses and final aspiration when studying the Swedish frequently occurring homonym “att”, namely that an “att” followed by a pause has an aspirated /t/. This could be a useful cue in developing algorithms for speech recognition and parsing. Although they point out that it would be fruitful to make a larger investigation, they suggest the following conclusion: “if “att” has final aspiration and is followed by a pause it is more likely that we are dealing with the realization of “att” as a subordinate conjunction, rather than the realization of “att” as an infinitive marker”.

Summary

In this paper we have looked through some important phonetic cues in English and Swedish signalling phrase boundary and also discourse finality, i.e. intensity, duration and fundamental frequency, with a closer look at special research studies regarding final lengthening, silent interval duration and final aspiration. Allthough there has been a substantial development of the field the last years, it is clear that automatic intepretation of fluent speech is a very complex task. There need to be further investigations made. For each language, it is also important to take combinations of different knowledge sources into consideration, when trying to signal or cancel out potential interpretations of communicative signals.

References

Berkovits, R. 1994. *Durational effects in final lengthening, gapping, and contrastive stress*. Language and Speech, vol. 37, pp. 237-250.

Blomberg, M., Elenius, K. 2000. *Automatisk igenkänning av tal*. Institutionen för tal, musik och hörsel, KTH. Stockholm.

Bruce, G., Granström, B., Gustafson, K., House, D. 1993. *Interaction of F0 and duration in the perception of prosodic phrasing in Swedish.* In Granström, B. & Nord, L. 1993 *Nordic Prosody VI,* 7-22. Almqvist & Wiksell, Stockholm.

Crystal, T.H. & House, A.S. 1990. *Articulation rate and the duration of syllables and stress groups in connected speech*. J. Acoust. Soc. Am. Vol 88, pp. 101-112.

Du Bois, J.W., Schuetze-Coburn, S., Cumming, S., Paolino, D. 1991. *Outline of discourse transcription*. In Edwards, J.A. & Lampert, M.D. 1993 *Talking Data Transcription and Coding in Discourse Research*. Lawrence Erlbaum Associates, Inc., Publishers. New Yersey.

Edwards, J.A. & Lampert, M.D. 1993 *Talking Data Transcription and Coding in Discourse Research*. Lawrence Erlbaum Associates, Inc., Publishers. New Yersey.

Fant, G. & Kruckenberg, A. 1989. *Preliminaries to the study of Swedish prose reading and reading style*. STL-QPSR 2.

Gold, B & Morgan, N. 2000. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music.* John Wiley & Sons, Inc. New York.

Horne, M. and Filipsson, M. 1995. *Computational modelling and generation of prosodic structure in swedish*. Department of Linguistics and Phonetics, University of Lund.

Horne, M., Strangert, E. & Heldner, M. 1995. *Prosodic boundary strength in Swedish: final lengthening and siletn interval duration.* Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, 170-173.

Igounet, S. 1998. *Système de reconnaissance automatique de la parole continue du français*. Thèse de l'École doctorale: Mathématiques et Informatique, l'Université d'Avignon et des pays de Vaucluse.

Johansson, V., Horne, M., Strömqvist, S. 2001. *Final aspiration as a phrase boundary cue in Swedish: the case of att "that"*. Department of Linguistics and Phonetics, Working Papers 49 (2001), 78–81. University of Lund.

Klatt, D.H. 1987. *Review of Text-to-Speech Conversion for English*. In Smithsonian Speech Synthesis History Project, Division of Information Technology and Society. National Museum Of American History, Smithsonian Institution - Washington, D.C. http://www.mindspring.com/~dmaxey/ssshp/dk_737a.htm, 2001-12-01.

Klatt, D.H. 1989. *Review of selected models of speech perception*. In Marslen-Wilson, W. *Lexical representation and process.* pp. 169-226. MIT Press, Cambridge.

McTear, M. F. 2001. *Spoken Dialogue Technology: Enabling the Conversational User Interface*. University of Ulster. Submitted to ACM Computing Surveys 2001.

Nakajima, S. & Allen, J. 1993. *A Study on Prosody and Discourse Structure in Cooperative Dialogues*. Rochester Tech Report No TRAINS-TN93-2, Sept. 1993.

Tannen, D. 1984. *Conversational style*. In Edwards, J.A. & Lampert, M.D. 1993 *Talking Data Transcription and Coding in Discourse Research*. Lawrence Erlbaum Associates, Inc., Publishers. New Yersey.